Should We Trust Occupational Income Scores?*

Martin Saavedra
Department of Economics
Oberlin College

Tate Twinam
Interdisciplinary Arts and Sciences
University of Washington Bothell

April 28, 2017

Abstract

Historical studies of labor market outcomes frequently suffer from a lack of data on individual income. The occupational income score (OCCSCORE) is often used as an alternative measure of labor market outcomes, particularly in studies of the U.S. prior to 1950. While researchers have acknowledged that this approach introduces measurement error, no effort has been made to quantify its impact on inferences. Using modern Census data, we find that the use of OCCSCORE biases results towards zero and can frequently result in statistically significant coefficients of the wrong sign. We show that a simple adjustment to OCCSCORE can substantially reduce this bias. We illustrate our results using the 1915 Iowa State Census, a rare source of pre-1950 earnings data. Using OCCSCORE in this context yields an attenuated wage gap for blacks and a statistically significant wage gap of the wrong sign for women; our adjusted OCCSCORE eliminates almost all of this bias. We also examine how bias due to the use of OCCSCORE affects estimates of intergenerational mobility using linked data from the 1850-1910 Censuses.

JEL codes: C21, C43, J71, N32
Keywords: OCCSCORE, occupational income score, non-classical measurement error, occupation, earnings gaps

*We are grateful for feedback from Brian Beach, Bruce Sacerdote, Randall Walsh, Mark Watson, Eugene White, and participants of the Northeast Ohio Economics Workshop and the Southern Economic Association meeting. Grant Goehring provided excellent research assistance. All errors are our own.

1 Introduction

Before 1940, data on individual wages and education are not available in the U.S. Census. Consequently, occupation is often the only measure of labor market outcomes available to economic historians. Occupation is a categorical variable; however, many economists use an index of occupational earnings potential as a continuous measure of historical labor market outcomes. One popular example is the 1950 occupational income score (OCCSCORE), which is the median income of an occupation in 1950. Occupational income scores have been used to examine earnings gaps going as far back at 1850, and studies using this approach have been published in numerous top journals in economics and other fields.[1]

Although occupational income scores should be correlated with earnings (for example, physicians and lawyers have higher occupational income scores than laborers), they are obviously an imperfect proxy for true earnings, and it is unclear how much bias this measurement error induces. Additionally, it is unclear if 1950 occupational income scores are good measures of labor market outcomes when examining Censuses several decades before 1950. While this potential bias has been acknowledged in the literature, no attempt has been made to quantify it and diagnose its impact on inferences. In this study, we attempt to measure this bias directly and examine how much it can be mitigated through adjustments to occupational income scores based on demographic and geographic variables universally available to economic historians.

We first develop a formal model of the measurement error problem posed by occupational income scores. The model allows us to determine when attenuation bias will occur and to explicitly quantify the magnitude of the bias. It also allows us to derive the condition under which the sign of an estimated coefficient will be incorrect. We then take this model to the data to estimate the OCCSCORE-induced bias. Because it is difficult to make historical data better, we analyze the performance of occupational income scores by making modern

[1] See section 2 for examples.

data worse.[2] We generate 2000-based occupational income scores and examine how well they predict income in the decades between 1950 and 2000. We then use this index to examine racial and gender earnings gaps from 1950 through 2000 and compare these to the true gaps estimated using actual earnings data. Finally, we show how the earnings gaps estimated using both parametrically and nonparametrically adjusted OCCSCOREs (based on race, gender, age, and region) compare to those generated using OCCSCORE and true earnings.

We find that although OCCSCORE is correlated with income even for censuses five decades removed from the base year, racial and gender earnings gaps are significantly attenuated when using OCCSCORE as a proxy for income. The use of OCCSCORE can result in statistically significant coefficients of the wrong sign up to 30 percent of the time in our modern data, particularly for variables indicating state of residency or state of birth.[3] This is even the case in earnings regressions where the sample is restricted to white males only. We find that adjusting OCCSCORE by race, gender, age, and region – an adjustment rarely made in empirical work–substantially reduces this bias.

To examine the performance of our adjusted OCCSCORE in a historical context, we exploit a rare source of pre-1950 earnings data: the 1915 Iowa State Census.[4] Estimated race and gender wage gaps in 1915 Iowa using true earnings data are sizable and negative; however, when using standard OCCSCORE as a proxy for earnings, the racial wage gap is attenuated by 80% and the gender wage gap is statistically significant and of the wrong sign. Both our parametrically and nonparametrically adjusted OCCSCOREs yield race and gender wage gaps very close to the true values. Finally, we conduct an analysis of OCCSCORE–induced bias in measures of intergenerational income transmission. This analysis is based

[2]Our approach is in the spirit of Romer (1986), who shows that excess volatility in unemployment time series during the pre-war era is an artifact of the interpolation methods used before the Current Population Survey. Applying the same interpolation methods to unemployment data during the post-war period results in similar levels of volatility.

[3]This finding is similar to that in Bertrand, Duflo and Mullainathan (2004) who find that difference-in-differences models that do not account for serial correlation in the error terms can result in statistically significant estimates of placebo treatment effects 40% of the time.

[4]The 1915 Iowa State Census data was digitized by Claudia Goldin and Lawrence Katz (Goldin and Katz, 2010).

on father–son pairs linked across the 1850, 1860, 1880, 1900, and 1910 decennial Censuses. We conclude with recommendations for future research in economic history.

2 Previous Literature

To understand how researchers use the occupational income score variable, we searched for papers containing either "OCCSCORE" or "Occupational Income Score" in top general interest journals and top field journals in labor economics and economic history. This search yielded the 20 papers listed in Table 1.[5] Most of the articles have been published within the last decade, with a median publication year of 2013.5. Fourteen use the log of occupational income score as a dependent variable, and consequently, we focus our empirical analysis on the log of occupational income score. Of these 20 papers, only three adjust occupational income scores by any demographic variables.[6] Typically, these papers analyze historical Census data for which income or wage data are not available and interpret occupational income score as a proxy for income. Some of the papers, however, present occupational income score along with wage/income data as an alternative measure of socioeconomic status (see Stephens and Yang (2014) or Chin (2005)). Some papers attempt to reduce the bias by limiting the sample to a particular demographic group, typically white males (see Bleakley (2010)). These papers often examine intergenerational mobility, racial and ethnic SES gaps, migrant selection, and the effects of schooling or health interventions.

This list underestimates how many researchers use occupational income scores or similar measures. Other papers in these journals may have used average or median income/wages by occupation as a dependent variable but do not refer to the variable as an occupational

[5] This search included articles in the following journals: *American Economic Journal: Applied Economics*; *American Economic Journal: Economic Policy*; *American Economic Review*; *Explorations in Economic History*; *Journal of Economic History*; *Journal of Human Resources*; *Journal of Labor Economics*; *Quarterly Journal of Economics*; and *Review of Economics and Statistics.* Other journals were searched, but yielded no results.

[6] The occupational earnings measure used by Collins and Wanamaker (2014) varies by race and region; Angrist (2002) varies by age and sex; and Collins (2000) presents results using both an unadjusted and a race-adjusted OCCSCORE.

Table 1: Published Studies using Occupational Income Scores

Article	Description	Adjusted	Log
Collins (2000)	Examines occupational mobility of blacks during the 1940s.	Both	Yes
Minns (2000)	Examines SES growth of immigrants relative to natives in the 1900 and 1910 censuses.	No	Yes
Angrist (2002)	Measures the effect of sex ratios of marriage markets and labor-force participation.	Yes	Yes
Chin (2005)	Estimates the long-run effects of adult incarceration in internment camps on labor-market outcomes.	No	No
Sacerdote (2005)	Measures the intergenerational effects of slavery.	No	No
Bleakley (2007)	Evaluates whether hookworm eradication during school-age affected human capital.	No	Yes
Bleakley and Lange (2009)	Tests the quantity-quality childbearing model using the eradication of hookworm as an exogenous shock to the returns for schooling.	No	Yes
Bleakley (2010)	Finds that childhood exposure to malaria decreases adult SES in Brazil, Colombia, Mexico, and the U.S.	No	Yes
Abramitzky, Boustan and Eriksson (2012)	Estimates the returns to migration and the degree of self-selection for Norwegian-U.S. migrants.	No	Yes
Lee (2013)	Examines how the repeal of Sunday closing laws affected year of schooling and adult outcomes.	No	Yes
Aaronson, Lange and Mazumder (2014)	Tests whether responses to the Rosenwald School initative are consistent with a quality-quantity childbearing model.	No	Yes
Collins and Wanamaker (2014)	Studies the returns to migration and self-selection for blacks during the Great Migration.	Yes	Yes
Cook, Logan and Parman (2014)	Finds no association between distinctively black names and socioeconomic status.	No	No
Stephens and Yang (2014)	Finds that prior estimates for the returns to schooling are not robust to region-specific birth year effects.	No	No
Collins and Wanamaker (2015)	Assesses self-selection of inter-regional and intra-regional migration.	No	Yes
Lleras-Muney and Shertzer (2015)	Analyzes the effect of English-only statutes on immigrant children literacy, years of schooling, and occupations.	No	Yes
Olivetti and Paserman (2015)	Creates pseudo-links using first names to estimate father-son and father-daughter elasticities for the intergenerational transmission of SES.	No	Yes
Saavedra (2015)	Estimates the effect of school-age incarceration in internment camps on adult outcomes.	No	No
Cook, Logan and Parman (2016)	Finds that 19th century blacks with distinctively black names live longer, suggesting cultural factors unrelated with SES affect mortality.	No	No
Massey (2016)	Assesses how the U.S. immigrant quota affected the selection of immigrants.	No	Yes

income score. Occupational income scores are also used in other fields, especially sociology. A Google Scholar search for research articles containing "OCCSCORE" or "occupational income score" currently yields 182 articles. Many of these articles are working papers that will eventually be published in top economics journals. For example, four NBER working papers in 2016 alone contain the phrase "OCCSCORE" or "Occupational Income Score."

3 Econometric Model

Suppose a researcher is interested in the following regression equation:

$$I_i = \beta_0 + \beta_1 Z_i + \boldsymbol{\beta}_2' \boldsymbol{X}_i + \delta_i \tag{1}$$

where I_i is individual i's income, Z_i is a scalar covariate of interest (such as race or sex), $\boldsymbol{X}_i$ is a vector of other covariates, and δ_i is an error term satisfying $\mathrm{E}\left[\, \delta \mid Z, \boldsymbol{X} \,\right] = 0$.[7] If data on I_i is available, estimation is straightforward. However, in many cases, the researcher does not observe income, but instead observes occupation. Suppose there are M occupations indexed by m. From a separate data source, the researcher also observes average or median earnings by occupation (perhaps in a different year). In place of I_i, the researcher assigns individual i from occupation m an occupational income score of $O_i = \mathrm{E}\left[\, I_i \mid \text{occupation}_i = m\right]$ or, more commonly, $O_i = \mathrm{Med}\left\{\, I_i \mid \text{occupation}_i = m\right\}$.[8] I_i is the desired outcome variable, but when I_i is unobserved, the occupational income score O_i must be used as a proxy. This results in the following model:

$$O_i = \beta_0 + \beta_1 Z_i + \boldsymbol{\beta}_2' \boldsymbol{X}_i + \underbrace{\delta_i - e_i}_{\epsilon_i} \tag{2}$$

The dependent variable in equation (2) suffers from measurement error $e_i = I_i - O_i$, which is now absorbed into the new error term ϵ_i along with the exogenous error term δ_i

[7] The isolation of Z from $\boldsymbol{X}$ is for illustrative purposes only; Z can be any element of the set of available covariates.

[8] Frequently, the natural logarithm of income or occupational income score is used instead, but this is immaterial for the following derivations.

from (1). Because $\mathrm{E}\left[\,\delta \mid Z, \boldsymbol{X}\,\right] = 0$, ordinary least squares estimation of equation (2) remains consistent so long as e_i is also uncorrelated with Z_i and $\boldsymbol{X}_i$. For this reason, measurement error in the dependent variable has received considerably less attention in graduate–level econometrics textbooks.[9]

However, in labor economics applications using occupational income scores, there is strong reason to believe that measurement error in the dependent variable is correlated with explanatory covariates. For example, indicator variables for race and sex are common explanatory variables. Racial minorities and females have historically earned less than white males even within the same occupation. Consequently, occupational income scores will overestimate earnings for these groups while underestimating earnings for white males, resulting in inconsistent OLS estimates. Although economic historians have acknowledged this bias, there has been no attempt to quantify it, and none of the alternative 1950–based occupational income scores in IPUMS account for any demographic or geographic factors (as can be seen in table 2). In this section, we analyze the nature of the bias theoretically, with an eye towards developing results that can be applied to real data in later sections.

Table 2: IPUMS Occupational Status Variables (1950 Base Year)

Variable name in IPUMS	Label	Basis of score	Source data
SEI	Duncan Socioeconomic Index	Income, Education, Prestige	1950 Census, 1947 North-Hatt Prestige Data
NPBOSS50	Nam-Powers-Boyd Occupational Status Score	Earnings, Education	1950 Census
PRESGL	Siegel Prestige Score	Prestige	1960s NORC Surveys
EDSCOR50	Occupational Education Score	Education	1950 Census
ERSCOR50	Occupational Earnings Score	Earnings	1950 Census
OCCSCORE	Occupational Income Score	Income	1950 Census

Notes: Adapted from "Table 2: Occupational Standing Variables included in IPUMS" found in the IPUMS User Guide.

Suppose the researcher is interested in the coefficient β_1 of Z, which may represent a gender earnings gap (if Z is an indicator for female), a particular racial earnings gap, or

[9] See, e.g., Wooldridge (2010) and Greene (2012).

any other earnings differential of interest. Naive OLS estimation of (2) yields the following asymptotic value for $\hat{\beta}_1$:

$$\operatorname{plim} \hat{\beta}_1 = \beta_1 - \frac{\operatorname{Cov}(e, \widetilde{Z})}{\operatorname{Var}(\widetilde{Z})} \tag{3}$$

where $\widetilde{Z}$ is the residual from the linear projection of Z onto X:[10]

$$\widetilde{Z}_i = Z_i - \boldsymbol{\gamma}' \boldsymbol{X}_i.$$

The rightmost term in (3) represents the bias due to measurement error in the dependent variable. Suppose that Z is an indicator for female; then, we would expect to find that $\beta_1 < 0$. If $\operatorname{Cov}(e, \widetilde{Z}) < 0$, i.e., women earn less than men within the same occupation (net of other factors), then measurement error leads to attenuation bias in $\hat{\beta}_1$. More generally, we expect that the correlation between $\widetilde{Z}$ and both income I and measurement error e has the same sign, i.e., if $\widetilde{Z}$ is associated with lower income overall, then it is likely associated with lower income within occupation. When this is the case, we have the following result:[11]

Proposition 3.1. *If* $\beta_1 > 0$ *and* $\operatorname{Cov}(e, \widetilde{Z}) > 0$*, then* $\beta_1 > \hat{\beta}_1$*. If* $\beta_1 < 0$ *and* $\operatorname{Cov}(e, \widetilde{Z}) < 0$*, then* $\beta_1 < \hat{\beta}_1$*.*

Proof. Proof of the first part of the proposition:

$$\begin{gathered} \operatorname{Cov}(e, \widetilde{Z}) > 0 \\ \implies \frac{\operatorname{Cov}(I, \widetilde{Z})}{\operatorname{Var}(\widetilde{Z})} > \frac{\operatorname{Cov}(I, \widetilde{Z})}{\operatorname{Var}(\widetilde{Z})} - \frac{\operatorname{Cov}(e, \widetilde{Z})}{\operatorname{Var}(\widetilde{Z})} \\ \implies \beta_1 > \hat{\beta}_1 \end{gathered}$$

which follows from the fact that

$$\beta_1 = \frac{\operatorname{Cov}(I, \widetilde{Z})}{\operatorname{Var}(\widetilde{Z})}$$

[10] $\boldsymbol{\gamma}$ is estimated from the linear regression $Z_i = \boldsymbol{\gamma}' \boldsymbol{X}_i + \eta_i$. $\widetilde{Z}$ represents the variation in Z that is orthogonal to the other regressors in (2).

[11] Henceforth, we use $\hat{\beta}_1$ to denote $\operatorname{plim} \hat{\beta}_1$, suppressing plim for notational clarity.

Reversing the inequalities yields the second part of the proposition. $\square$

This measurement error will frequently result in attenuation bias while preserving the correct sign of the coefficient. However, there are cases where a sign reversal can occur. The following result develops the necessary and sufficient condition for this sign reversal in the above cases:

Proposition 3.2. *Assume* $\beta_1 \neq 0$ *and*

$$\operatorname{sgn}(\operatorname{Cov}(I, \widetilde{Z})) = \operatorname{sgn}(\operatorname{Cov}(e, \widetilde{Z}))$$

holds. Then, $\operatorname{sgn}(\beta_1) = \operatorname{sgn}(\hat{\beta}_1)$ *if and only if*

$$\frac{\operatorname{Corr}(I, \widetilde{Z})}{\operatorname{Corr}(e, \widetilde{Z})} > \frac{\operatorname{Var}(e)}{\operatorname{Var}(I)}$$

Proof. Case 1: Assume that $\operatorname{Cov}(I, \widetilde{Z}) > 0$ and $\operatorname{Cov}(e, \widetilde{Z}) > 0$. Then, $\beta_1 > 0$ and

$$\begin{aligned} \hat{\beta}_1 > 0 &\iff \frac{\operatorname{Cov}(I, \widetilde{Z})}{\operatorname{Var}(\widetilde{Z})} - \frac{\operatorname{Cov}(e, \widetilde{Z})}{\operatorname{Var}(\widetilde{Z})} > 0 \\ &\iff \frac{\operatorname{Cov}(I, \widetilde{Z})}{\operatorname{Var}(\widetilde{Z}) \operatorname{Var}(I) \operatorname{Var}(e)} > \frac{\operatorname{Cov}(e, \widetilde{Z})}{\operatorname{Var}(\widetilde{Z}) \operatorname{Var}(I) \operatorname{Var}(e)} \\ &\iff \frac{\operatorname{Corr}(I, \widetilde{Z})}{\operatorname{Corr}(e, \widetilde{Z})} > \frac{\operatorname{Var}(e)}{\operatorname{Var}(I)} \end{aligned}$$

Case 2: Assume that $\operatorname{Cov}(I, \widetilde{Z}) < 0$ and $\operatorname{Cov}(e, \widetilde{Z}) < 0$. Then, $\beta_1 < 0$ and

$$\begin{aligned} \hat{\beta}_1 < 0 &\iff \frac{\operatorname{Cov}(I, \widetilde{Z})}{\operatorname{Var}(\widetilde{Z})} - \frac{\operatorname{Cov}(e, \widetilde{Z})}{\operatorname{Var}(\widetilde{Z})} < 0 \\ &\iff \frac{\operatorname{Cov}(I, \widetilde{Z})}{\operatorname{Var}(\widetilde{Z}) \operatorname{Var}(I) \operatorname{Var}(e)} < \frac{\operatorname{Cov}(e, \widetilde{Z})}{\operatorname{Var}(\widetilde{Z}) \operatorname{Var}(I) \operatorname{Var}(e)} \\ &\iff \frac{\operatorname{Corr}(I, \widetilde{Z})}{\operatorname{Corr}(e, \widetilde{Z})} > \frac{\operatorname{Var}(e)}{\operatorname{Var}(I)} \end{aligned}$$

$\square$

This result shows that sign reversal may occur when the correlation between $\widetilde{Z}$ and income is small relative to the correlation between $\widetilde{Z}$ and the measurement error, and gives an exact condition for when this should take place. Along with the previous result, we have characterized the expected bias resulting from measurement error due to the use of occupational income scores. While the theory gives some insight into the expected direction of the bias, the magnitude and importance of the bias is not obvious. The next section brings the theory to the data to explicitly compute the magnitude of the bias and sign reversals seen in some commonly estimated regressions.

What can be done about this measurement error bias? Clearly it can be reduced by minimizing $\mathrm{Cov}(e, \widetilde{Z})$ to the extent possible. Fortunately, there is a straightforward solution available using historical data. The most common covariates include age and indicators for race/ethnicity, sex, and some geographic breakdown.[12] All of these are likely to be correlated with e. We can reduce this correlation by constructing an adjusted version of the occupational income score which stratifies on these common demographic variables. Below, we consider two alternative measures. The first is a nonparametrically adjusted OCCSCORE, which reflects the median income for a given occupation and base year within cells defined by sex, age, race, and region:

$$Np\_Adj\_O_i = \mathrm{Med}\left\{ I_i \mid \mathrm{occupation}_i = m,\, sex_i,\, age_i,\, race_i,\, region_i \right\}. \tag{4}$$

The advantage of this measure is that it allows for arbitrary interactions between all of the adjustment variables. The disadvantage is that stratifying on so many variables may result in small or empty cells, leading to excessively variable or missing occupational income scores for some individuals.

An alternative that avoids this problem involves a parametric strategy. For example,

[12] Another common demographic variable that could be used is an indicator for foreign–born status. However, we caution that this may be misleading. The composition of the foreign–born population in the US in 1950 differs substantially from that in earlier years such as 1900 and 1850 (in both racial/ethnic terms and human capital terms). Thus, this adjustment may lead to inaccurate results.

one could estimate the following regression of income in a given base year on a series of occupation ($\boldsymbol{O}$), sex (S), age ($\boldsymbol{A}$), race ($\boldsymbol{R}$), and geographic region ($\boldsymbol{G}$) indicator variables:

$$I_i = \beta_0 + \boldsymbol{\beta}_1^{'}\boldsymbol{O}_i + \beta_2 S_i + \boldsymbol{\beta}_3^{'}\boldsymbol{A}_i + \boldsymbol{\beta}_4^{'}\boldsymbol{R}_i + \boldsymbol{\beta}_5^{'}\boldsymbol{G}_i + \epsilon_i. \tag{5}$$

The fitted coefficients can be used to generate an adjusted OCCSCORE for each possible individual as follows:

$$\begin{aligned} P\_Adj\_O_i &= \mathrm{E}\left[\, I_i \mid \text{occupation}_i = m,\ sex_i,\ age_i,\ race_i,\ region_i\right] \\ &= \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1^{'}\boldsymbol{O}_i + \hat{\beta}_2 S_i + \hat{\boldsymbol{\beta}}_3^{'}\boldsymbol{A}_i + \hat{\boldsymbol{\beta}}_4^{'}\boldsymbol{R}_i + \hat{\boldsymbol{\beta}}_5^{'}\boldsymbol{G}_i. \end{aligned} \tag{6}$$

This strategy does not allow for interaction effects, but it is computationally simple and generates an adjusted OCCSCORE for every type of individual. In most of the analysis below, we focus on the more general nonparametrically adjusted OCCSCORE; however, we do compare its performance with that of the parametric adjusted OCCSCORE in both modern and historical contexts and find that the results tend to be very similar.[13] Unless explicitly stated otherwise, references to "adjusted OCCSCORE" below refer to the nonparametrically adjusted OCCSCORE $Np\_Adj\_O_i$.

In the following sections, we examine the performance of our adjusted OCCSCORE measures relative to the standard OCCSCORE in contexts where true income data is available to provide a baseline. We find that our alternative measures substantially reduce the sizable attenuation bias induced by measurement error when occupation is used as a proxy for income.

[13]There are a wide range of semiparametric adjusted OCCSCOREs one could consider. We examined the performance of a specification allowing average income to vary with age in a quadratic fashion while still stratifying on occupation, sex, race, and region. On the Iowa sample examined in section 5.1, it performed similarly to both the parametric and nonparametric adjusted OCCSCOREs.

4 Results

4.1 Persistence of Occupational Income

Our empirical analysis uses data from the 1950-2000 Censuses downloaded from the Integrated Public Use Microdata Series (IPUMS) published by the Minnesota Population Center (MPC). The most commonly used occupational income score is the OCCSCORE variable, which is a weighted average of the median earnings for males and females for each occupational category in 1950. The median earnings data come from a 3.3% sample of the 1950 Census and excludes individuals without positive income. In this section, we (1) replicate the 1950 OCCSCORE variable using the 1% sample of the 1950 Census, (2) test how well 1950 OCCSCORE predicts median earnings in future census years, (3) construct a 2000-based OCCSCORE using the same method as the MPC, and (4) test how well the 2000-based OCCSCORE predicts median earnings from past censuses. If the 2000-based OCCSCORE successfully proxies for median earnings in 1950, then the 1950 OCCSCORE may be a reasonable proxy for median occupational earnings in 1900.

Figure 1 presents scatter plots of 1950 OCCSCORE against median earnings from the 1950-2000 censuses. The size of each circle corresponds to the number of individuals in each occupational category during that census year. Panel A of Table 3 regresses 1950 OCCSCORE on median earnings for each census year and weights each regression by the size of occupation cell. Regressing 1950 OCCSCORE on median income yields an R^2 of 0.96 and slope coefficient of 1.08. The rest of the panels/columns show the persistence of occupational income scores from 1960 to 2000. Median income appears to remain strongly correlated with 1950 occupational income scores. Predicting 1950 OCCSCORE using median income from the 1960-2000 Censuses yields R^2's between 0.79 and 0.95. The farther away from the 1950 one gets, the less predictive power median income has on occupational income score.

Figure 2 replicates this exercise for the 2000-based occupational income score, and the

Table 3: Predicting OCCSCORE using median income

Panel A: Predicting 1950-based OCCSCORE using median income

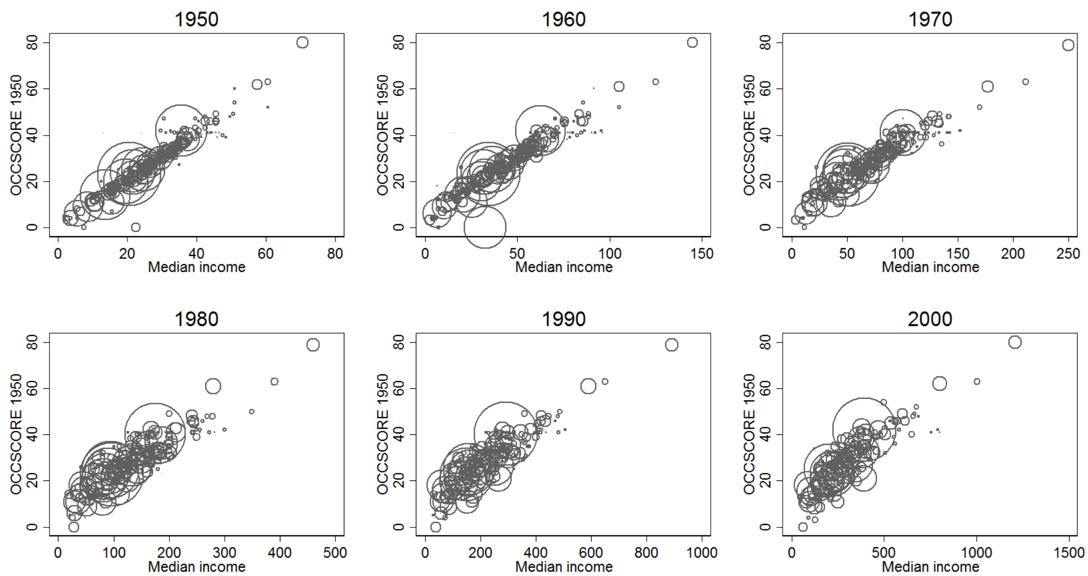
	Census year					
	1950	1960	1970	1980	1990	2000
	(1)	(2)	(3)	(4)	(5)	(6)
Median income	1.088***	0.546***	0.312***	0.162***	0.0882***	0.0676***
	(0.0456)	(0.0230)	(0.0129)	(0.00926)	(0.00653)	(0.00583)
Constant	-0.712	3.868***	6.305***	7.324***	9.085***	8.659***
	(0.854)	(0.843)	(0.925)	(0.931)	(0.960)	(1.164)
N	268	269	258	220	219	187
R^2	0.965	0.948	0.909	0.842	0.812	0.786

Panel B: Predicting 2000-based OCCSCORE using median income

	Census year					
	1950	1960	1970	1980	1990	2000
	(1)	(2)	(3)	(4)	(5)	(6)
Median income	11.40***	6.250***	4.083***	2.306***	1.330***	1.000***
	(1.351)	(0.482)	(0.202)	(0.103)	(0.0175)	(0.000)
Constant	-7.986	17.32	14.77	1.708	10.84*	0.000
	(35.80)	(21.02)	(12.53)	(11.59)	(4.237)	(0.000)
N	187	187	186	185	185	187
R^2	0.728	0.824	0.899	0.920	0.978	1

Notes: Data are from the 1% sample of the U.S. Census downloaded from IPUMS. Each column regresses OCCSCORE on median earnings in that Census year. The dependent variable in the IPUMS 1950 OCCSCORE for Panel A, and the constructed 2000-based OCCSCORE for Panel B. The coefficient for 1950 is not equal to 1 in Panel A because OCCSCORE comes from a Census report derived from a 3.3% sample of the U.S. Census. The unit of observation is an occupation, and each occupation is weighted by the number of people in that occupation.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 1: 1950 occupational income score and median income

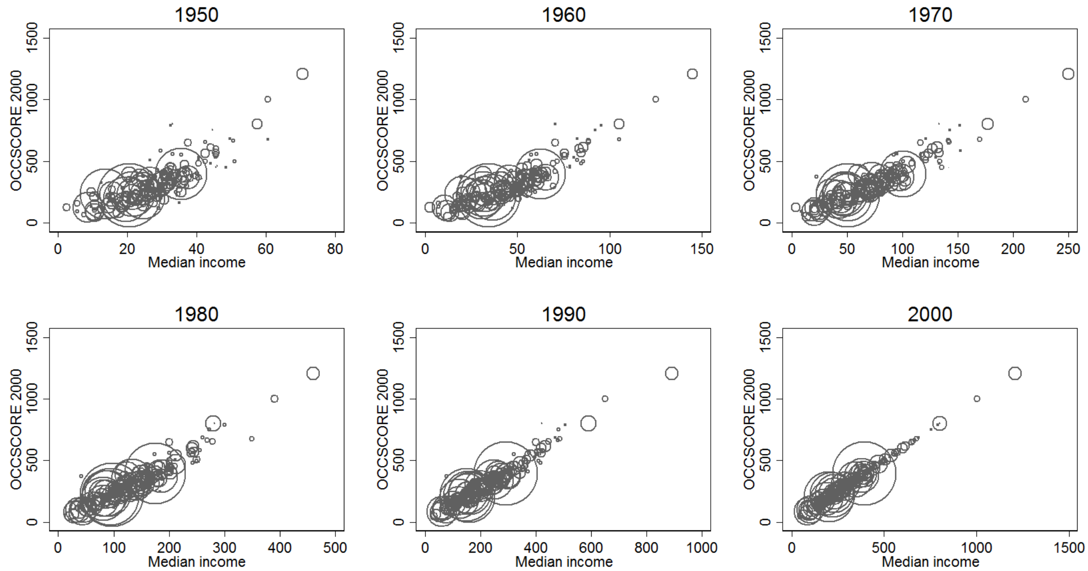


Notes: Median earnings for each occupation are from the 1% sample of the U.S. census. The 1950-based OCCSCORE are from IPUMS and are a weighted average median male and female earnings in 1950 for each occupation from a 3.3% sample. The size of each circle corresponds to the number of individuals in the occupational category during that census year.

regression results are in Panel B of Table 3. Since median earnings and the 2000-based OCCSCORE come from the same sample, the $R^2 = 1$ for 2000 with a slope coefficient of 1. For each decade removed from 2000, the R^2 decreases, implying that OCCSCORE is becoming worse as a proxy for median earnings. Even 50 years removed from the base year, however, the $R^2 = 0.72$, implying that OCCSCORE is a strong proxy for median earnings.

The previous results suggests that the rank of occupational earnings is relatively stable over time. Whether this trend continues to Censuses before 1950 remains an open and, without additional earnings data, unanswerable question. If there were large income shocks to particular occupations, we would likely see changes in the occupational distribution as workers left occupations with negative earnings shocks and entered occupations with positive earnings shocks. Consequently, if we observe the same level of occupational reshuffling between Census years in the 1900 to 1950 period as we observe from 1950 to 2000, it is likely that shocks to occupational earnings are occurring at a similar rate. To measure occupational

Figure 2: 2000 occupational income score and median income



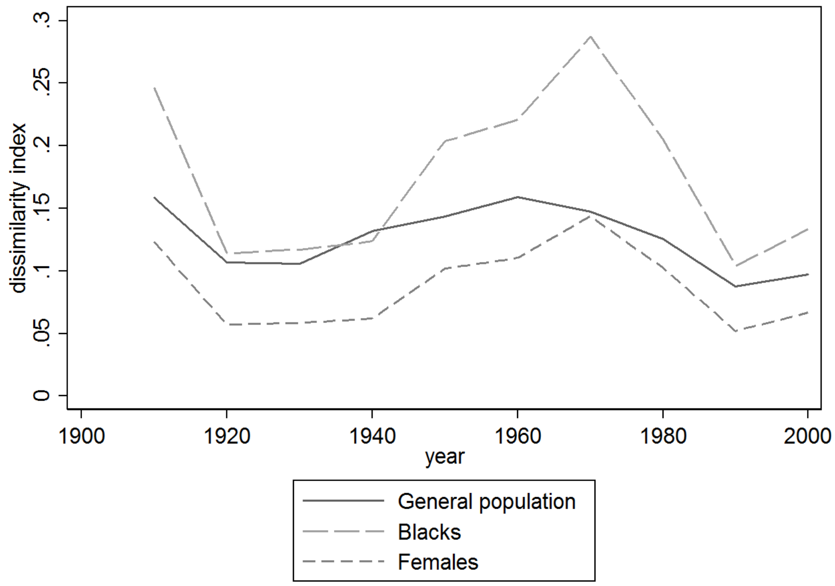
Notes: Median earnings for each occupation are from the 1% sample of the U.S. census. The 2000-based OCCSCORE are the median earnings from the 2000 census for individuals in each occupation. The size of each circle corresponds to the number of individuals in the occupational category during that census year.

reshuffling, we calculate an occupational dissimilarity index between adjacent census years:

$$\text{dissimilarity} = \frac{1}{2}\sum_{i=1}^{N}\left|\frac{\text{occ}_{i,t}}{\text{pop}_t} - \frac{\text{occ}_{i,t-10}}{\text{pop}_{t-10}}\right| \tag{7}$$

where N is the number of occupational categories, $\text{occ}_{i,t}$ is the number of individuals in occupation i in census year t, and pop_t is the number of individuals with any occupation in year t. The variables indexed $t-10$ are from the previous census year. The dissimilarity index measures the proportion of people in one census year who would have to change their occupation to equalize the occupational distribution to that of the previous census year. The results of this exercise are graphed in Figure 3. For the general population, the occupational dissimilarity index is relatively constant at approximately 10% during the 1910-2000 period. Females reshuffle occupations at a slightly lower rate than the general population. Blacks have reshuffled occupations at a higher rate, with the occupational dissimilarity index increasing during the civil rights movement, peaking in 1970, and subsequently returning to

Figure 3: Occupational Dissimilarity Index between Adjacent Census Years



Notes: Data are from the IPUMS. The dissimilarity index measures the proportion of the population that would have to switch occupations to equalize the occupational distribution from the previous census year.

levels close to the general population. Because the rate of occupational reshuffling in the general population has been relatively constant from 1910 to 2000, we expect that changes in occupational earnings between 1910 and 1950 occurred at a similar rate as changes between 1950 and 2000.

4.2 Errors of Magnitude

In this section, we analyze the magnitude of bias induced when using OCCSCORE and adjusted OCCSCORE as a proxy for income in an earnings regression. For the moment, we focus on estimating racial and gender earnings gaps. Gelman and Tuerlinckx (2000) and Gelman and Carlin (2014) introduce the Type M error rate as the expected value of an estimate divided by the true parameter value, conditional on the estimate being statistically different from zero. In this context, the true earnings gap is the earnings gap found using actual income data and the estimated earnings gap is the gap using a proxy for income

(either OCCSCORE or adjusted OCCSCORE).[14]

Table 4 reports results from three regression models. The first panel regresses the log of earnings on a set of dummies for state of residence, sex, race, and nativity. In addition to these dummy variables, the regression includes age and age squared. We run the regressions separately for every Census year from 1950 to 2000. Because we assume researchers would have used earnings instead occupational income scores if earnings data were available, we treat these coefficients as the true parameters that researchers would like to estimate. The second panel runs the same regressions as the first panel, but instead of log earnings as the dependent variable, the dependent variable is the log of the 2000-based OCCSCORE. The dependent variable for the last panel is the log of the adjusted 2000-based OCCSCORE. The adjusted OCCSCORE is non-parametrically adjusted and measures the median earnings in 2000 for each occupation, sex, race, and region cell. We restrict the sample to adults ages 25-65 who were in the labor force.

As expected, earnings gaps for women and blacks decline over time, and this is reflected in all three models. However, the magnitude of the gap is highly attenuated when the log of OCCSCORE is used in place of true earnings. For all years, the coefficients for sex, race, and nativity are of the same sign in all panels, but the coefficients in panels B and C suffer from substantial attenuation bias. Adjusting OCCSCORE by race, sex, age, and region greatly reduces this bias, but does not completely eliminate it.

Figure 4 graphs the implied earnings gaps using the three measures. The earnings gap estimated using adjusted OCCSCORE more closely mirrors the true earnings gap than unadjusted OCCSCORE. In addition to the non-parametically adjusted OCCSCORE, we graph estimated earnings gaps using the parametrically adjusted OCCSCORE. The parametrically adjusted OCCSCORE comes from the predicted values of a regression of earnings on a set of dummies for race, sex, age, and region (without interaction terms) using only 2000 Census data. Since the parametrically adjusted and non-parametrically adjusted OCCSCORE

[14] Although we do not formally take the expectation of the earnings gaps, the large sample size of the Census ensures that the standard errors are small and the estimated coefficients will be close to their expectations.

Table 4: Comparing Models with Earnings, 2000 OCCSCORE, and Adjusted 2000 OCCSCORE

	(1)	(2)	(3)	(4)	(5)	(6)
Census Year:	1950	1960	1970	1980	1990	2000
Dependent variable: log of earnings						
Female	-0.569***	-0.760***	-0.823***	-0.757***	-0.613***	-0.501***
	(0.00522)	(0.00254)	(0.00215)	(0.00200)	(0.00184)	(0.00169)
Black	-0.470***	-0.530***	-0.373***	-0.235***	-0.276***	-0.247***
	(0.00927)	(0.00469)	(0.00393)	(0.00355)	(0.00350)	(0.00300)
Other race	-0.508***	-0.363***	-0.181***	-0.0218**	-0.249***	-0.192***
	(0.0524)	(0.0199)	(0.0139)	(0.00841)	(0.00498)	(0.00332)
U.S. born	0.0613***	0.116***	0.123***	0.186***	0.182***	0.188***
	(0.00779)	(0.00463)	(0.00435)	(0.00393)	(0.00366)	(0.00297)
Dependent variable: log of 2000 OCCSCORE						
Female	-0.203***	-0.250***	-0.258***	-0.244***	-0.194***	-0.162***
	(0.00256)	(0.00123)	(0.00112)	(0.000981)	(0.000890)	(0.000832)
Black	-0.277***	-0.276***	-0.243***	-0.187***	-0.188***	-0.155***
	(0.00455)	(0.00228)	(0.00205)	(0.00174)	(0.00169)	(0.00147)
Other race	-0.323***	-0.134***	-0.0484***	0.00457	-0.125***	-0.0860***
	(0.0257)	(0.00964)	(0.00724)	(0.00413)	(0.00241)	(0.00163)
U.S. born	0.0724***	0.0796***	0.0747***	0.0966***	0.0893***	0.0983***
	(0.00382)	(0.00225)	(0.00227)	(0.00193)	(0.00177)	(0.00146)
Dependent variable: log of adjusted 2000 OCCSCORE						
Female	-0.473***	-0.517***	-0.525***	-0.491***	-0.459***	-0.446***
	(0.00297)	(0.00141)	(0.00125)	(0.00108)	(0.000975)	(0.000941)
Black	-0.395***	-0.380***	-0.316***	-0.253***	-0.249***	-0.219***
	(0.00527)	(0.00260)	(0.00228)	(0.00192)	(0.00185)	(0.00167)
Other race	-0.469***	-0.245***	-0.179***	-0.110***	-0.289***	-0.240***
	(0.0298)	(0.0110)	(0.00805)	(0.00454)	(0.00264)	(0.00184)
U.S. born	0.0562***	0.0769***	0.0771***	0.104***	0.0856***	0.105***
	(0.00443)	(0.00256)	(0.00252)	(0.00212)	(0.00194)	(0.00165)
N	117241	456799	543841	696142	888106	1086343

Notes: Data are from the 1% sample of the U.S. Census downloaded from IPUMS. Each regression also controls for state of residency, age, and age squared. The 2000-based OCCSCORE is the median earnings of an occupation in the 2000 Census. The adjusted OCCSCORE is median earnings in the 2000 Census for every race, sex, age, and region cell. Standard errors are in parentheses.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

perform similarly, we focus on non-parametrically adjusted OCCSCORE for the remainder of this analysis.

Our estimates of the female/male earnings ratio are similar to the extant literature (Goldin, 1990, pp. 62). The female/male earnings ratio declined between 1950-1960, after which the gender gap slowly narrowed. Margo (2016) provides census estimates of the black/white earnings gap that are similar to ours. Black income increased relative to whites during the 1960s and 1970s, but the ratio has not narrowed significantly since the 1980s. Smith (1984) estimates the black/white income gap by assigning each individual the average income of race, sex, and age cell from the 1970 census. These estimates, produced at least a decade before the IPUMS OCCSCORE variable was regularly in use, are in essence an adjusted OCCSCORE.

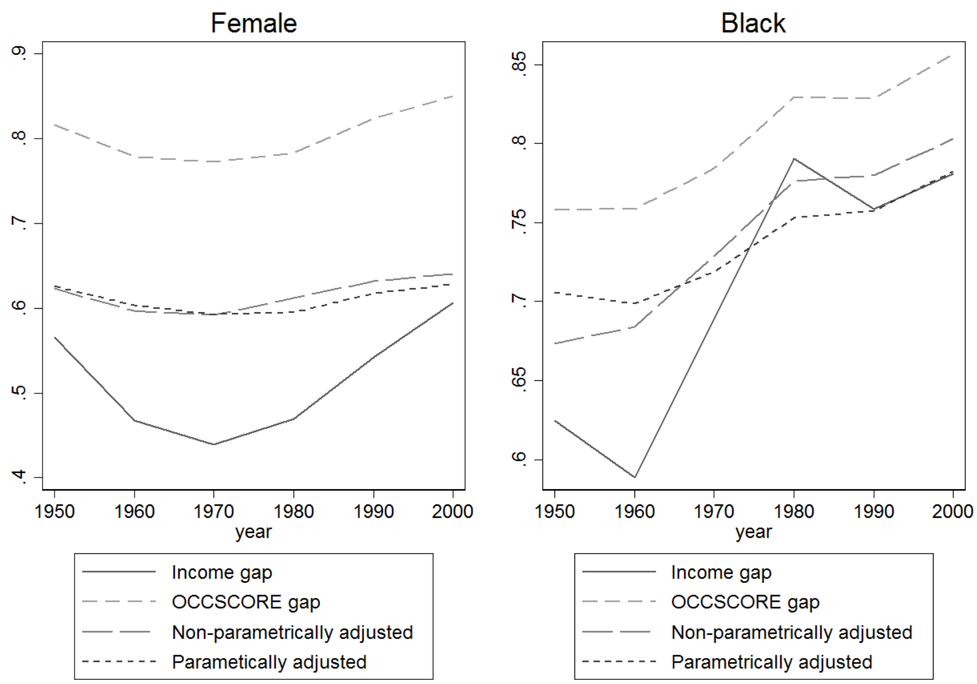
4.3 Errors of Sign

If researchers are primarily concerned with the direction of an effect as opposed to its magnitude, the previous results suggest that qualitative conclusions may not be seriously affected by the use of occupational income scores.[15] The use of OCCSCORE as the dependent variable did not result in sign changes for gender and racial wage differences. This result does not generalize to regressors with signs less predictable than race and gender indicators. Here, we show that OCCSCORE frequently results in errors of sign, or Type S errors. A Type S error occurs when the true population parameter is non-zero and the estimate is statistically significant and of the wrong sign (see Gelman and Carlin (2014) and Gelman and Tuerlinckx (2000)).

In this section, we consider four models. We run a regression with 185 dummy explanatory variables: a set of dummy variables for state of residence, age, race, birthplace, farm status, family size, marital status, number of families in the household, and relationship to the household head. Then we run the model with a 2000-based occupational income score as

[15] However, see section 5.1 below, where the use of OCCSCORE does result in a sign change for the gender wage gap which is statistically significant.

Figure 4: Earnings ratios using earnings, OCCSCORE, and adjusted OCCSCORE



Notes: The data are from IPUMS Ruggles, Genadek, Goeken, Grover and Sobek (2015). The graph displays the implied female/male and black/white income ratios from Table 4. Note that the gaps are conditional on age, age squared, a dummy variable for US-born, and state of residency. The female earnings gap is conditional on race, and the black/white earnings gap is conditional on sex. OCCSCORE uses a 2000-based occupational income score, whereas adjusted OCCSCORE is a 2000-based occupational income score conditional on race, sex, age, and region. The sample is restricted to those between ages 25 and 65 who were in the labor force.

Table 5: How often the coefficients conflict with the "true model"

Model	1950	1960	1970	1980	1990	2000
OCCSCORE	.286	.173	.200	.119	.065	.060
OCCSCORE (clustered s.e.)	.054	.076	.059	.038	.016	.016
Adjusted OCCSCORE	.097	.049	.022	.022	.016	.022
Adjusted OCCSCORE (clustered)	.032	.016	.005	.016	.005	.011

Notes: Data are from the 1% samples of the U.S. Census downloaded from IPUMS. We regress the measure of labor market outcomes on 185 dummy variables for state of residence, age, race, birthplace, farm status, family size, marital status, number of families in the household, and relationship to the household head. The "true model" uses log of earnings. Each cell displays the proportion of those estimates that are statistically significant in both models and of the wrong sign.

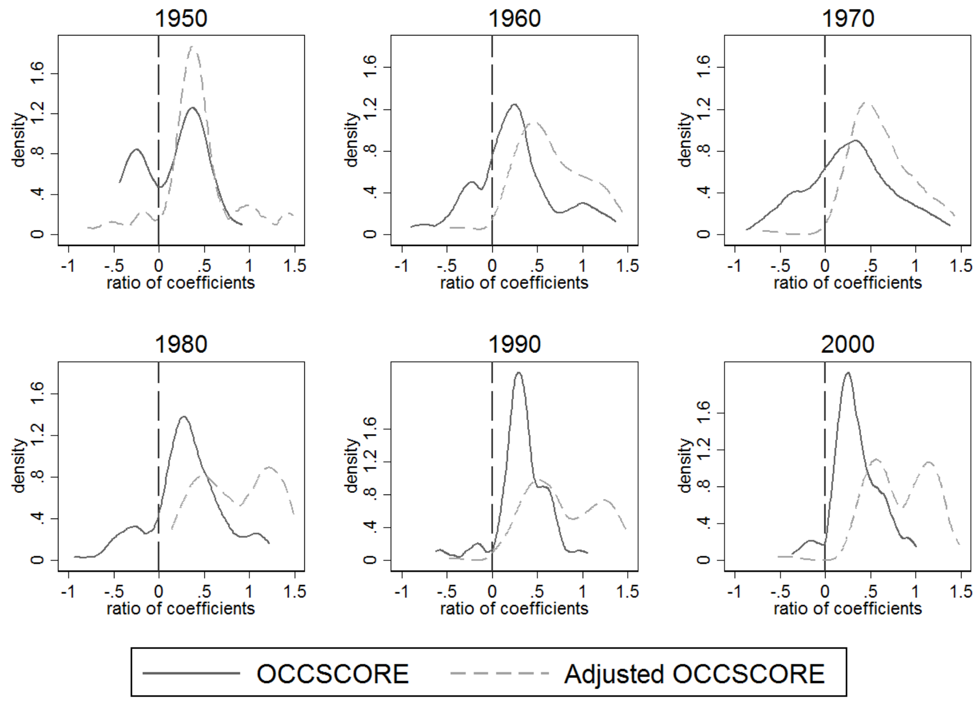
the dependent variable and then again with a 2000-based occupational income score that is adjusted for race, sex, age, and region.

We run both of these models using the default standard errors and standard errors clustered at the occupation level. The intuition for clustering at the occupation level is as follows: suppose the dependent variable is a 2000-based occupational income score and we are running a regression using 1950 data. Suppose a particular occupation earned less in 1950 than in 2000 relative to the rest of the population. In this case, the measurement error for all individuals of that occupation in 1950 will be correlated.

We then compare how often these models give conflicting results compared to the "true model" in which the dependent variable is log income. Many researchers are satisfied to use estimators that are biased towards zero since the sign of the estimator is likely to be the same as that of the parameter they are trying to estimate. A more serious problem occurs when researchers find spurious results that are statistically significant and of the wrong sign. For this reason, we say that the two models conflict for a particular coefficient if they produce opposite signs, but the model using occupational income score is statistically significant.

The results from this exercise are shown in Table 5. The results suggest that out of the 185 coefficients, up to 29 percent of them are statistically significant and of the wrong sign. The problem is worse for years far away from the base year and is greatly reduced by clustering the standard errors at the occupation level and by adjusting the occupational income score

Figure 5: Density of Ratios of Estimated to True Coefficients using OCCSCORE and Adjusted OCCSCORE



Notes: Data are from IPUMS Ruggles et al. (2015). The figures display the density of the ratio of the estimated coefficients (using a 2000-based OCCSCORE) and the "true" coefficient using observable income. Each year contains 185 regression coefficients and includes a set of dummy variables for state of residence, age, race, birthplace, farm status, family size, marital status, number of families in the household, and relationship to the household head. The sample is restricted to those between ages 25 and 65 who were in the labor force.

by race, sex, age, and region. The problem is almost eliminated when doing both. Figure 5 graphs kernel density estimates of the ratio of the estimated and true coefficients. The density for adjusted OCCSCORE is closer to being centered around one (less attenuation bias) and has less weight to the left of zero (conflicting signs). Using OCCSCORE leads to less accurate results the farther away from the base year we get, whereas estimates from adjusted OCCSCORE are likely to be of the same sign even 50 years prior to the base year.

Many papers attempt reduce the bias of OCCSCORE by restricting historical samples to only white males. Women were not nearly as prominent in the labor market historically as they are today, and it is unclear whether blacks were paid close to median earnings

within occupation. Studies from the antebellum era also limit their samples to whites since the majority of Southern blacks were slaves during the 1850 and 1860 censuses. Figure 6 repeats the previous exercise while restricting the sample to white males. The 2000-based OCCSCORE uses data from both females and non-whites, but the regression coefficients are estimated using only white males. This is analogous to a researcher using the default IPUMS OCCSCORE (which is constructed using data from all races and both genders) but then estimating regressions while restricting the sample to white males. Although restricting the sample to white males does reduce the probability of a type S error, once one is three decades removed from the base year, the probability of a type S error is significantly less using adjusted OCCSCORE. Type M errors are also smaller using adjusted OCCSCOREs, as the densities are closer to being centered around one.

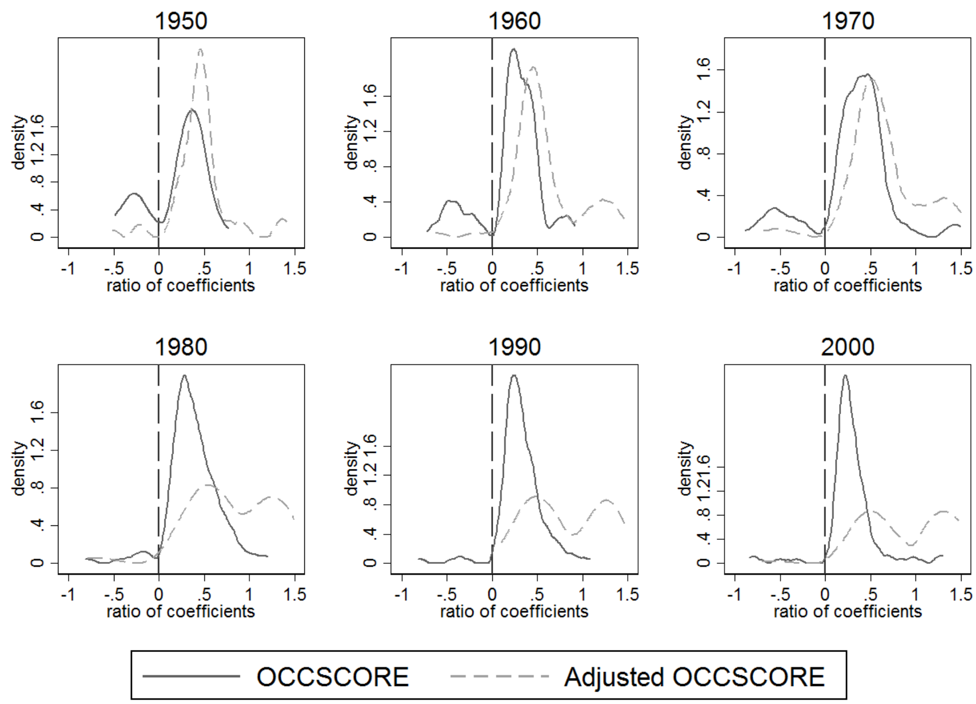
5 Applications

5.1 Earnings Gaps in the Iowa State Census, 1915

The analysis in section 4 shows that when examining gender and racial wage gaps, adjusted OCCSCOREs substantially outperform the standard OCCSCORE using modern decennial Census data. To assess the extent to which this conclusion applies in a more historical context, we exploit a rare source of pre-1950 income data, the 1915 Iowa State Census (Goldin and Katz, 2010).[16] This was the first Census in the US to collect data on income prior to 1940. The sample digitized by Goldin and Katz (2010) contains records on 5.5% of the urban population drawn from three of Iowa's largest cities: Des Moines, Dubuque, and Davenport. It also contains 1.8% of the population of counties not containing a major city; the ten counties used span the geography of the state. We compare racial and gender wage gaps estimated using standard OCCSCOREs and our two adjusted OCCSCOREs to those estimated using true earnings.

[16]This data has also been used to examine intergenerational mobility by Feigenbaum (2017).

Figure 6: Density of Ratios of Estimated to True Coefficients using OCCSCORE and Adjusted OCCSCORE: White Males Only



Notes: Data are from IPUMS Ruggles et al. (2015). The entire population is used to construct 2000-based OCCSCOREs and adjusted OCCSCOREs, but only white males are used in the regression. The figures display the density of the ratio of the estimated coefficients (using a 2000-based OCCSCORE) and the "true" coefficient using observable income. Each year contains 185 regression coefficients, and includes a set of dummy variables for state of residence, age, race, birthplace, farm status, family size, marital status, number of families in the household, and relationship to the household head. The sample is restricted to those between ages 25 and 65 who were in the labor force.

For the estimation, we restrict the sample to those between the ages of 25 and 65 and exclude those with missing occupation data or zero/missing earnings.[17] The census reports occupation categories according to the 1940 scheme. We cross-walked these with the 1950 scheme to match individuals in 1915 to their 1950 OCCSCORE.[18] For some individuals, we are not able to compute our nonparametric adjusted OCCSCORE due to empty cells in the 1950 Census sample; in the estimation below, we only include individuals for which we have all three OCCSCORE measures for comparability. The final sample includes 11,707 individuals. We estimate the wage gap between between whites and blacks and men and women; approximately 1% of sample is black (122 obs) and 13% of sample is female (1,515 obs).

Table 6: Earnings Gaps in the 1915 Iowa State Census

	Log of earnings	Log of 1950 OCCSCORE	Log of param. adj. 1950 OCCSCORE	Log of nonparam. adj. 1950 OCCSCORE
	(1)	(2)	(3)	(4)
Black	-0.445***	-0.086***	-0.379***	-0.355***
	(0.0609)	(0.0319)	(0.0322)	(0.0394)
Female	-0.453***	0.021**	-0.367***	-0.473***
	(0.0186)	(0.0097)	(0.0098)	(0.0120)
Age	0.046***	0.012***	0.051***	0.046***
	(0.0045)	(0.0024)	(0.0024)	(0.0029)
Age^2	-0.000***	-0.000***	-0.001***	-0.001***
	(0.0001)	(0.0000)	(0.0000)	(0.0000)
Estimation	OLS	OLS	OLS	OLS
Observations	11,707	11,707	11,707	11,707
R^2	0.063	0.004	0.144	0.138

Notes: Linear regressions of earnings measures on blacks and female indicators as well as a quadratic polynomial in age. Sample excludes those whose race is recorded as Missing, Mixed, or Asian (24 observations), those who are below the age of 25 or above the age of 65, those with missing occupation data, and those with zero or missing earnings. Sample is further restricted to individuals for which all OCCSCORE measures could be calculated. *** p<0.01, ** p<0.05, * p<0.1

[17] We also exclude those whose race is recorded as missing; those whose race is recorded as Mixed or Asian (24 observations) are also excluded because the sample is too small to reliably estimate wage gaps for these groups.

[18] In some cases, the 1940 scheme aggregated some occupations; for example, bookkeepers, accountants, and cashiers fall into one occupation category in 1940 but are disaggregated into three separate categories in 1950. There are 7 occupation categories in 1940 (out of 194 total) that cannot be matched uniquely to a 1950 occupation; we exclude individuals falling into these categories.

In column (1) of table 6, we report the coefficients from a regression of log earnings on an indicator for black and female as well as a quadratic polynomial for age. Women and blacks earn substantially less than white men. As is typical, earnings increase with age but at a diminishing rate. In column (2), we replace log earnings with the log of the standard 1950 OCCSCORE. The results change substantially: the black-white wage gap coefficient declines by 80%, and the gender wage gap is now positive and statistically significant. The use of standard 1950 OCCSCOREs as a proxy for earnings leads to misleading results in this context.

Moving to column (3), we replace the unadjusted OCCSCORE outcome with the log of our parametrically adjusted 1950 OCCSCORE. The results change dramatically: the black and female wage gaps are now large, negative, and similar in magnitude to those computed using log earnings as the outcome variable. The impact of age is now substantially larger as well and very similar in magnitude to the estimate from column (1). Column (4) repeats this exercise using the log of nonparametric adjusted 1950 OCCSCORE. The results are similar to those in columns (1) and (3), indicating that both adjusted OCCSCORE measures perform similarly to each other and serve as excellent proxies for true earnings.

5.2 Estimates of Intergenerational Mobility

Labor economists often measure intergenerational mobility by regressing a son's socioeconomic status on his father's socioeconomic status:

$$I_i^{\text{son}} = \beta_0 + \beta_1 I_i^{\text{father}} + \delta_i \tag{8}$$

where I_i^{son} is the log income of a son observed during adulthood, and I_i^{father} is the log income of a father observed while the son was a child. The transmission coefficient β_1 is an elasticity typically between 0 and 1, with 1 representing perfect immobility between generations and 0 representing perfect mobility. Historical evidence on occupational mobility

across generations relies heavily on occupational incomes scores instead of income for two reasons. First, to obtain data on father's and son's labor market outcomes in the Census, one needs to link across census years, which is typically only possible using given and surnames. Names do not become publicly available in the census until 72 years after the census year, meaning occupations are the only available labor market outcomes for both fathers and sons. Second, estimates of how intergenerational mobility have changed over time require data spanning at least three generations, implying that such estimates must make use of historical data.

Let $e_i^{\text{son}} = I_i^{\text{son}} - O_i^{\text{son}}$ and $e_i^{\text{father}} = I_i^{\text{father}} - O_i^{\text{father}}$ be the measurement error from using an occupational index (either OCCSCORE or adjusted OCCSCORE) for the son and father, respectively. Then researchers estimate:

$$O_i^{\text{son}} = \beta_0 + \beta_1 O_i^{\text{father}} + \underbrace{\beta_1 e_i^{\text{father}} - e_i^{\text{son}} + \delta_i}_{\epsilon_i}. \tag{9}$$

This regression differs from the model in Section 3 since OCCSCORE appears on both the left-hand and right-hand side of the regression. The measurement errors e_i are likely to be smaller if one uses a demographically adjusted OCCSCORE instead of an unadjusted OCCSCORE, since racial, age, and region differences in occupational earnings will not be captured in e_i. However, when using unadjusted OCCSCORE, some of the measurement error is likely to cancel out since e_i^{son} is positively correlated with e_i^{father}. Our estimate of the transmission coefficient will be biased if $Cov(O_i^{\text{father}}, \beta_1 e_i^{\text{father}} - e_i^{\text{son}}) \neq 0$. If there is little intergenerational mobility, in which case β_1 is close to 1, and if the son's measurement error is highly correlated with the father's measurement error, then the second term of covariance is close to zero. Alternatively, suppose $e_i^{\text{son}} = \tilde{\beta}_0 + \tilde{\beta}_1 e_i^{father} + \nu_i$, where ν_i is independent of all other variables. The transmission coefficient $\tilde{\beta}_1$ reflects that fathers who earn above average within their occupations are likely to have sons who earn above average within occupations. Then, $Cov(O_i^{\text{father}}, \beta_1 e_i^{\text{father}} - e_i^{\text{son}}) = Cov(O_i^{\text{father}}, \beta e_i^{\text{father}} - \tilde{\beta}_1 e_i^{\text{father}})$.

Thus, the bias from estimating equation 6 using OCCSCOREs will be small so long as the transmission in overall income from father to son is similar to the transmission of excess income within occupation. For these reasons, using occupational income scores instead of income should lead to smaller bias when estimating intergenerational mobility compared to estimating racial and gender income gaps, and estimates of intergenerational mobility using adjusted and unadjusted OCCSCORE should be similar.

In Table 7, we provide estimates of intergenerational mobility using the IPUMS linked data sets. These data link the 1% samples of the 1850, 1860, 1900, and 1910 Censuses to the 1880 complete count. The sample is restricted to those who during the first census year were children of the household head, male, and no older than 15 years old. We regress the log of a son's OCCSCORE (during the second census year) on the log of the father's OCCSCORE (during the first census year), and then repeat the regression using adjusted OCCSCORE. The resulting coefficients are elasticities, with higher coefficients implying occupational immobility. Row 1 of columns (2) and (3) of table 7 are replications of the estimates in row 7 of table 3 of Olivetti and Paserman (2015). In column (2), we find identical results to Olivetti and Paserman (2015), and in column (3) we find similar results. We present estimates for whites using all samples, and for blacks using samples in which both the father and son are observed in the postbellum period.

For whites, we find greater occupational mobility using adjusted OCCSCORE relative to OCCSCORE for the 1860-1880, 1880-1900, and 1880-1910 samples. For the 1850-1880 sample of whites, we find similar occupational mobility. We find that there is less occupational mobility for blacks between 1880-1900, perhaps because before the Great Migration most Southern-born blacks remained in the South, and adjusted OCCSCORE varies by region. We find similar levels of occupational mobility using OCCSCORE and adjusted OCCSCORE for blacks between 1880-1910. Although the coefficients using adjusted and unadjusted OCCSCORE are statistically different, they are all of the same sign and of roughly similar magnitudes.

Table 7: Estimates of Intergenerational Mobility

Dependent variable: log of son's OCCSCORE						
Sample	whites	whites	whites	whites	blacks	blacks
	1850-1880	1860-1880	1880-1900	1880-1910	1880-1900	1880-1910
log of father's OCCSCORE	0.393***	0.452***	0.511***	0.417***	0.125***	0.172***
	(0.0217)	(0.0170)	(0.0119)	(0.0128)	(0.0447)	(0.0565)
N	2849	4049	9146	7776	630	400
Dependent variable: log of son's adjusted OCCSCORE						
Sample	whites	whites	whites	whites	blacks	blacks
	1850-1880	1860-1880	1880-1900	1880-1910	1880-1900	1880-1910
log of father's adjusted OCCSCORE	0.405***	0.289***	0.356***	0.368***	0.242***	0.166**
	(0.0188)	(0.0179)	(0.0131)	(0.0128)	(0.0452)	(0.0663)
N	2319	3308	7324	6124	455	273

Notes: Data are from the IPUMS linked data files. These data are from 1% samples of the 1850, 1860, 1900, and 1910 Censuses linked to the 1880 complete count. The sample is restricted to those who during the first census year were children of the household head, male, and no older than 15 years old. Standard errors are in parentheses. * $p < 0.1$; ** $p < 0.05$; *** $p < 0.01$.

6 Conclusion

Using modern Census data, we find that even though occupational income is highly correlated over time, occupational income scores systematically underestimate racial and gender income gaps by a substantial margin. Furthermore, other standard earnings regression covariates like state of residency and state of birth indicators can be of the wrong sign up to 30 percent of the time. Adjusting occupational income scores by race, gender, age, and region (variables available in every Census going back to 1850) significantly reduces the attenuation bias and limits the probability that the estimates are of the wrong sign. We also show that earnings regressions using OCCSCORE can be misleading even when the sample is restricted solely to white males.

To examine the performance of adjusted OCCSCORE in a historical context, we examine data from the 1915 Iowa State Census, which collected data on both occupation and earnings. We find that estimated race and gender wage gaps in 1915 Iowa using true earnings data are sizable and negative; however, when using standard OCCSCORE as a proxy for earnings, the racial wage gap is attenuated by 80% and the gender wage gap is actually positive and statistically significant. Both our parametrically and nonparametrically adjusted OCCSCOREs yield race and gender wage gaps very close to the true values. We also

conduct an analysis of OCCSCORE–induced bias in measures of intergenerational income transmission. This analysis is based on father–son pairs linked across the 1850, 1860, 1880, 1900, and 1910 decennial Censuses. While the standard OCCSCORE does not perform as poorly here (due to correlated errors), we nonetheless find improvements when employing an adjusted OCCSCORE. Our results strongly suggest that future research in economic history and other fields should utilize adjusted occupational income scores whenever possible.

References

Aaronson, Daniel, Fabian Lange, and Bhashkar Mazumder (2014). Fertility transitions along the extensive and intensive margins. *American Economic Review*, **104**(11), pp. 3701–3724.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson (2012). Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration. *American Economic Review*, **102**(5), pp. 1832–1856.

Angrist, Josh (2002). How do sex ratios affect marriage and labor markets? Evidence from America's second generation. *Quarterly Journal of Economics*, pp. 997–1038.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan (2004). How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics*, **119**(1), pp. 249–275.

Bleakley, Hoyt (2007). Disease and development: evidence from hookworm eradication in the American South. *Quarterly Journal of Economics*, **122**(1), p. 73.

——— (2010). Malaria eradication in the Americas: A retrospective analysis of childhood exposure. *American Economic Journal: Applied Economics*, **2**(2), pp. 1–45.

Bleakley, Hoyt and Fabian Lange (2009). Chronic disease burden and the interaction of education, fertility, and growth. *Review of Economics and Statistics*, **91**(1), pp. 52–65.

Chin, Aimee (2005). Long-Run Labor Market Effects of Japanese American Internment during World War II on Working-Age Male Internees. *Journal of Labor Economics*, **23**(3), pp. 491–525.

Collins, William J (2000). African-American economic mobility in the 1940s: a portrait from the Palmer Survey. *Journal of Economic History*, **60**(03), pp. 756–781.

Collins, William J and Marianne H Wanamaker (2014). Selection and economic gains in the great migration of african americans: New evidence from linked census data. *American Economic Journal: Applied Economics*, **6**(1), pp. 220–252.

——— (2015). The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants. *Journal of Economic History*, **75**(04), pp. 947–992.

Cook, Lisa D, Trevon D Logan, and John M Parman (2014). Distinctively black names in the American past. *Explorations in Economic History*, **53**, pp. 64–82.

——— (2016). The mortality consequences of distinctively black names. *Explorations in Economic History*, **59**, pp. 114–125.

Feigenbaum, James J. (2017). Multiple Measures of Historical Intergenerational Mobility: Iowa 1915 to 1940. *Economic Journal*.

Gelman, Andrew and John Carlin (2014). Beyond power calculations assessing type S (sign) and type M (magnitude) errors. *Perspectives on Psychological Science*, **9**(6), pp. 641–651.

Gelman, Andrew and Francis Tuerlinckx (2000). Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, **15**(3), pp. 373–390.

Goldin, Claudia (1990). *The gender gap: An economic history of American women*. New York: Cambridge University Press.

Goldin, Claudia and Lawrence Katz (2010). The 1915 Iowa State Census Project: ICPSR28501-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research, `http://doi.org/10.3886/ICPSR28501.v1`.

Greene, William H. (2012). *Econometric analysis*. Prentice Hall, 7th edition.

Lee, Dara N (2013). The impact of repealing Sunday closing laws on educational attainment. *Journal of Human Resources*, **48**(2), pp. 286–310.

Lleras-Muney, Adriana and Allison Shertzer (2015). Did the Americanization Movement Succeed? An Evaluation of the Effect of English-Only and Compulsory Schooling Laws on Immigrants. *American Economic Journal: Economic Policy*, **7**(3), pp. 258–290.

Margo, Robert A (2016). Obama, Katrina, and the Persistence of Racial Inequality. *The Journal of Economic History*, **76**(02), pp. 301–341.

Massey, Catherine G (2016). Immigration quotas and immigrant selection. *Explorations in Economic History*, **60**, pp. 21–40.

Minns, Chris (2000). Income, cohort effects, and occupational mobility: a new look at immigration to the United States at the turn of the 20th century. *Explorations in Economic History*, **37**(4), pp. 326–350.

Olivetti, Claudia and M Daniele Paserman (2015). In the Name of the Son (and the Daughter): Intergenerational Mobility in the United States, 1850–1940. *American Economic Review*, **105**(8), pp. 2695–2724.

Romer, Christina (1986). Spurious volatility in historical unemployment data. *The Journal of Political Economy*, pp. 1–37.

Ruggles, Steven, Katie Genadek, Ronald Goeken, Josiah Grover, and Matthew Sobek (2015). Integrated Public Use Microdata Series Version 6.0, Machine-readable database. Minneapolis: University of Minnesota.

Saavedra, Martin (2015). School quality and educational attainment: Japanese American internment as a natural experiment. *Explorations in Economic History*, **57**, pp. 59–78.

Sacerdote, Bruce (2005). Slavery and the intergenerational transmission of human capital. *Review of Economics and Statistics*, **87**(2), pp. 217–234.

Smith, James P (1984). Race and human capital. *American Economic Review*, **74**(4), pp. 685–698.

Stephens, Melvin and Dou-Yan Yang (2014). Compulsory education and the benefits of schooling. *American Economic Review*, **104**(6), pp. 1777–1792.

Wooldridge, Jeffrey M. (2010). *Econometric analysis of cross section and panel data*. MIT Press, 2nd edition.